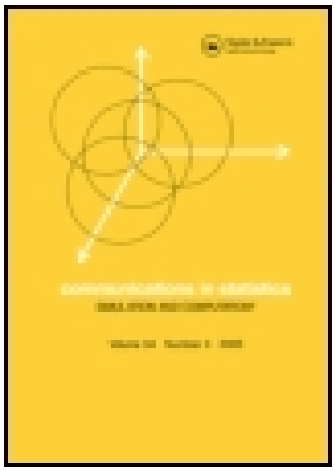

Communications in Statistics - Simulation and Computation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

The combined model: A tool for simulating correlated counts with overdispersion

George Kalema[ab], Samuel Iddi[ac] & Geert Molenberghs[ad]

[a] I-Biostat, KU Leuven, Kapucijnenvoer 35, B3000, Leuven, Belgium

[b] School of Statistics and Applied Economics, Makerere University, P.O. Box 7062, Kampala, Uganda

[c] Department of Statistics, University of Ghana, P.O. Box LG 115, Legon-Accra, Ghana

[d] I-Biostat, Universiteit Hasselt, Martelarenlaan 42, 3500, Hasselt, Belgium

Accepted author version posted online: 05 Aug 2014.

To cite this article: George Kalema, Samuel Iddi & Geert Molenberghs (2014): The combined model: A tool for simulating correlated counts with overdispersion, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2014.906610

To link to this article: http://dx.doi.org/10.1080/03610918.2014.906610

The combined model: A tool for simulating correlated counts with overdispersion

George KALEMA[1,2], **Samuel IDDI**[1,3], **Geert MOLENBERGHS**[1,4]

[1] *I-Biostat, KU Leuven, Kapucijnenvoer 35, B3000 Leuven, Belgium*

[2] *School of Statistics and Applied Economics, Makerere University, P.O. Box 7062, Kampala, Uganda*

[3] *Department of Statistics, University of Ghana, P.O. Box LG 115, Legon-Accra, Ghana*

[4] *I-Biostat, Universiteit Hasselt, Martelarenlaan 42, 3500 Hasselt, Belgium*

Abstract

The combined model as introduced by Molenberghs et al. (2007, 2010) has been shown to be an appealing tool for modeling not only correlated or overdispersed data but also for data that exhibit both these features. Unlike techniques available in the literature prior to the combined model, which use a single random-effects vector to capture correlation and/or overdispersion, the combined model allows for the correlation and overdispersion features to be modeled by two sets of random effects. In the context of count data, for example, the combined model naturally reduces to the Poisson-normal model, an instance of the generalized linear mixed model in the absence of overdispersion and it also reduces to the negative-binomial model in the absence of correlation. Here, a Poisson model is specified as the parent distribution of the data conditional on a normally distributed random effect at the subject or cluster level and/or a gamma distribution at observation level. Importantly, the development of the combined model and surrounding derivations have relevance well beyond mere data analysis. It so happens that the combined model can also be used to simulate correlated data. If a researcher is interested in comparing marginal models via Monte Carlo simulations, a necessity to generate suitable correlated count data arises. One option is to induce correlation via random effects but calculation of such quantities as the bias is then not straightforward. Since overdispersion and correlation are simultaneous features of longitudinal count data, the combined model presents an appealing

framework for generating data to evaluate statistical properties, through a pre-specification of the desired marginal mean (possibly in terms of the covariates and marginal parameters) and a marginal variance-covariance structure. By comparing the marginal mean and variance of the combined model to the desired or pre-specified marginal mean and variance, respectively, the implied hierarchical parameters and the variance-covariance matrices of the normal and Gamma random effects are then derived from which correlated Poisson data are generated. We explore data generation when a random intercept or random intercept and slope model is specified to induce correlation. The data generator, however, allows for any dimension of the random effects although an increase in the random-effects dimension increases the sensitivity of the derived random effects variance-covariance matrix to deviations from positive-definiteness. A simulation study is carried out for the random-intercept model and for the random intercept and slope model, with or without the normal and Gamma random effects. We also pay specific attention to the case of serial correlation.

Key Words: Copulas, Correlated data, Multivariate Gamma distribution, Poisson distribution.

1 Introduction

Research today generates a lot of data that have to be analyzed and summarized into meaningful and informative statements. Analysis is done using statistical methods that depend on the kind of data at hand. In medical research, it is often the case that data on a patient is profiled longitudinally in the sense that each patient is followed repeatedly or observed at multiple points over time. This introduces the phenomenon of correlated data because observations from one patient will be more related or similar than observations across different patients. A lot of research has already been committed to the analysis of correlated data. For example, Molenberghs and Verbeke (2005) and Verbeke and Molenberghs (2000) focus on methods for the analysis of discrete

and continuous longitudinal data, respectively. In the context of continuous or normal longitudinal data, calculations are computationally easier than in the non-normal case because the model for the response variable given random effects is the normal distribution and that of the random effects is the normal distribution as well. The two combined and integrating over the random effects leads to a normal distribution as the marginal model. In the non-normal case though, the model for the outcome variable and the random effects combined does not lead, in general, to closed-form solutions for the marginal model. Even if it does, expressions tend to be cumbersome. This is due to the lack of the elegant and convenient multivariate distributions analogous to the case of longitudinal data that can be assumed normally distributed. This poses computational and interpretational challenges. Specific to count data, which is of interest here, evaluation of the multivariate Poisson distribution grows in computational complexity with an increase in the dimensions due to the summations inherent in the distribution (Karlis, 2003). It is therefore of interest to find alternative means of analysis of correlated count data. One alternative is the generalized linear mixed model (GLMM) proposed by Breslow and Clayton (1993). This model accounts for the correlation by use of effects specific to a subject or study unit (random effects) and then derives the marginal distribution as a result of combining a random-effects distribution with a Poisson distribution for the data given the random effects. Molenberghs et al. (2007, 2010) have introduced the so-called combined model (CM) as a tool to model data that is not only correlated but also overdispersed. Overdispersion may occur when the model restricts the data in the sense that the variance expected from the model is less than that observed in the data. It is commonly encountered in data assumed to follow a binomial distribution, correlated or uncorrelated, correlated Bernoulli/binary random variables, correlated or independent observations arising from counting processes (Poisson data), and time-to-event/survival data. This is due to the mean-variance relationship inherent in the distributions that are assumed to be the data generating mechanisms. Overdispersion is, however, not an issue in the case of independent Bernoulli observations. Research has shown overdispersion to be caused by, for example, missing covariates and the presence of correlation between individual

responses or clustering, among others. Depending on outcome type and model, not accounting for overdispersion may lead to bias in some or all parameters; it definitely biases precision estimates. The result is then usually smaller *p*-values for the statistical tests as well as, of course, confidence intervals that are narrower than should be if overdispersion were properly handled. This means that inference based on such statistical analyses is questionable and may be misleading.

Solutions have been proposed in the literature and implemented in statistical software to account for overdispersion. The negative-binomial (NEGBIN) model for count data is one such tool which assumes the count data to have the Poisson as the parent distribution and a Gamma distribution for the extra parameter that accounts for overdispersion. The resulting marginal distribution is then the negative-binomial distribution. Note that earlier statistical analyses were generally only able to account for either correlation or overdispersion, but not both. But, given data that exhibit both features, it is a necessity to account for both in analyses, indeed. We refer to Section 2 for a detailed description of the GLMM, negative-binomial, and combined models.

We now turn to data-generation, the aspect which this paper contends to contribute to. It is common practice in statistics to carry out Monte-Carlo (MC) simulations in which samples are randomly drawn from probability distributions to mimick statistical processes that can be used to study properties of statistical methods. Simulation of correlated Poisson random variables is a topic of ongoing research and various methods have been proposed in the literature to this end, some of which include: the overlapping sums (Madsen and Dalthorp, 2007; Mardia, 1970; Kocherlakota and Kocherlakota, 1992, 2001); Lognormal-Poisson hierachy; Normal to Anything (NorTA; Cario and Nelson, 1997, 1998; Nelson, 2006; Mardia, 1970; Li and Hammond, 1975), and extensions thereof (Yahav and Shmueli, 2012; Ghosh and Pasupathy, 2012; Shin and Pasupathy, 2007; Avramidis et al., 2009; Park and Shin, 1998; Downer and Moser, 2001). See also Devroye (1986) for an overview on random variate generation. These tools yield correlated Poisson random variables with the specification of the Poisson means and the desired or target correlation structure. Most

of these methods, however, suffer from such limitations as: severe computational restrictions; difficulty achieving the target correlation; generated variables are required to be overdispersed; low correlations obtained; correlations constrained to be strictly positive; etc. Another approach is to use random effects to induce the correlation, thereby generating data from a hierarchical model. If the simulation is in the context of hierarchical models, this approach would be fine. However, whenever interest is in population-averaged or marginal models, the parameters used in the hierarchical model do not have a 1:1 correspondence with those in the marginal model. Given such a tool as the combined model that incorporates the two common features of count data, namely, overdispersion and correlation, it certainly is essential to generate data from such a method whenever interest is in simultaneously investigating these features. In this paper, we present the combined model as a tool to generate correlated Poisson random variables.

The rest of the paper is organized as follows. Section 2 reviews the modeling background. In Section 3, the focus is on data generation. A simulation study is set up in Section 4; results are presented in Section 5.

2 Overview of the models

2.1 Notation

Our focus in this paper is the generation of correlated count or Poisson random variables for K independent subjects in a study with subject i having measurements Y_{ij}, $i = 1, \ldots, K$, $j = 1, \ldots, n_i$. This is based on specification of the mean model in terms of an $n_i \times p$ known design matrix X_i, a p-dimensional fixed-effects parameter vector $\boldsymbol{\beta}$ and Z_i, an $n_i \times q$ design matrix for the random effects of subject i.

2.2 Modeling discrete correlated data

In dealing with discrete univariate data, generalized linear models (GLM; Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989; Agresti, 2002), which are a class of fixed-effects models unifying linear, logistic, and Poisson regression models among others, is the standard approach for analysis. The GLM generalizes the linear regression model in that the linear component, expressed in terms of covariates, relates to the response variable via a link function. In the presence of correlation, an extension of the GLM framework to the so-called generalized linear mixed model (GLMM; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Molenberghs and Verbeke, 2005) is commonly used. The GLMM modifies the linear predictor in the GLM to include unknown subject-specific effects in addition to the fixed effects. These subject-specific effects or random effects are, in practice, usually assumed to follow a normal distribution for reasons of convenience and availability of software, but any other distribution could be used in principle. Specific to count data, the mathematical expression of the GLMM is

$$\begin{aligned} Y_{ij}|\mathbf{b}_i &\sim \text{Poi}(\lambda_{ij}), \\ \ln(\lambda_{ij}) &= X_{ij}^{\top}\boldsymbol{\beta} + Z_{ij}^{\top}\mathbf{b}_i, \\ \mathbf{b}_i &\sim N(0, D), \end{aligned} \tag{1}$$

whereby the conditional distribution of the observations from a subject i given the random effects $\mathbf{b}_i$ is Poisson with a rate parameter λ_{ij} that is log-linearly related to covariates. Fitting these models is done by maximizing the marginal likelihood resulting from integrating (3) over the random effects. Closed form expressions for these integrals do not exist in all cases but Molenberghs et al. (2007,

2010) derived the marginal mean and covariance for the Poisson case as

$$\mu_{ij} = \ln(\lambda_{ij}) = X_{ij}^{\top}\boldsymbol{\beta} + 0.5Z_{ij}^{\top}DZ_{ij}, \tag{2a}$$

$$\text{var}(\mathbf{Y}_i) = M_i + M_i(e^{Z_iDZ_i^{\top}} - J_i)M_i, \tag{2b}$$

respectively, where J_i is a matrix of 1's and M_i is a diagonal matrix with entries μ_{ij}. Also, the higher-order marginal moments and the marginal joint distribution can be derived in closed form for the Poisson case (Molenberghs et al. 2010).

2.3 Modeling overdispersion

As mentioned in the introduction, overdispersion is a phenomenon where the observed variance in the data is greater than what is expected or predicted by the model. An obvious check for overdispersion is to compare the sample mean and sample variance. It is expected that the mean and variance are the same for the Poisson case, and deviations from this point to the more rarely encountered case of underdispersion (the observed sample variance is less than the predicted or expected model variance) or overdispersion. Indeed, models that account for overdispersion have been proposed in the literature and even implemented in statistical software packages like SAS and R, for example. Some references in this light are Hinde and Demétrio (1998a, 1998b), Breslow (1984), Lawless (1987), and Molenberghs and Verbeke (2005). In dealing with overdispersed data, one way forward is to assume a two-stage approach for the response such that in stage 1, a distribution is considered for the response or outcome variable, given a random effect $f(y_i|\mathbf{b}_i)$, and in stage 2, a model for the random effects $f(\mathbf{b}_i)$ is specified. Combining the two stages and integrating over the random effects results in the marginal model:

$$f(y_i) = \int f(y_i|\mathbf{b}_i)f(\mathbf{b}_i)d\mathbf{b}_i. \tag{3}$$

For count data, frequently the assumption that $Y_i|\lambda_i \sim \text{Poi}(\lambda_i)$ is made, together with allowing λ_i to be a random variable assumed to follow a Gamma distribution with $E(\lambda_i) = \mu_i$ and $\text{var}(\lambda_i) = \sigma_i^2$. The marginal distribution is then the negative-binomial distribution. Extension to the case of correlated or hierarchical count data is rather easy as shown in Section 3.2 of Molenberghs et al. (2010).

2.4 Modeling correlation and overdispersion

Analysis of data with both correlation and overdispersion features is a continuing area of research, indeed. The introduction of the combined model by Booth et al. (2003) and Molenberghs et al. (2007, 2010) quite flexibly accounts for these features simultaneously. Please note that it is not our intention to present a comprehensive literature review of the combined model and its associates. Rather, we reflect on the combined model as a data generator but refer to, for example, Winkelmann (2004, 2008), Sutradhar (2011), Chid and Quddus (2003), Deb and Holmes (2000) and related references therein for discussions of similar approaches and further details on this matter. The combined model brings together the two models discussed in Sections 2.2 and 2.3 in the presence of both correlation and overdispersion. It also reduces to the GLMM in the presence of correlation and overdispersion as far as described by the normal random effects, or the negative-binomial model in the presence of overdispersion but not correlation. The CM is given by

$$\begin{aligned} Y_{ij} &\sim \text{Poi}(\lambda_{ij}^*), \\ \lambda_{ij}^* = \theta_{ij}\lambda_{ij} &= \theta_{ij}\exp(\boldsymbol{x}_{ij}^\top\boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top\mathbf{b}_i), \\ \boldsymbol{\theta}_i &\sim \text{Gamma}(\text{mean} = \mathbf{1}, \text{variance} = \Sigma_i), \\ \mathbf{b}_i &\sim N(0, D), \end{aligned} \tag{4}$$

where θ_{ij}, the entries in $\boldsymbol{\theta}_i$, are the overdispersion parameters introduced at observation level. If we assume the θ_{ij} to be independent as is often done in practice, then the association is only induced by the $\mathbf{b}_i$ and the θ_{ij} would cover the overdispersion not accounted for by the normal random effects. Then, Σ_i is reduced to a diagonal matrix. Alternatively, the θ_{ij} can be allowed to be correlated as well such that Σ_i can take on more general structures. This implies the use of some form of Multivariate Gamma (MGamma) distribution. For example, Σ_i can be chosen such that there is a time-dependence, or other covariate dependencies, in the association structure. Evidently, as is also the case in the linear mixed model, when random effects and general Σ_i are present, the user needs to carefully ensure that the resulting marginal model is identifiable. A classical counterexample from the linear mixed model setting is a random intercept combined with a compound-symmetry residual structure. This leads to fully aliased parameters. The marginal mean and the marginal variance-covariance matrix take the form:

$$\begin{aligned} E(Y_{ij}) = \mu_{ij} &= \theta_{ij} \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + 0.5 \boldsymbol{z}_{ij}^\top D \boldsymbol{z}_{ij}), \\ \text{var}(\mathbf{Y}_i) &= M_i + M_i (P_i - \mathbf{J}_i) M_i, \end{aligned} \tag{5}$$

where $M_i = \text{diag}(\boldsymbol{\mu}_i)$ and

$$P_i = e^{\left(0.5 Z_i D Z_i^\top\right)} (\Sigma_i + J_i)\, e^{\left(0.5 Z_i D Z_i^\top\right)}.$$

Here, J_i is a matrix of ones. Note that we make use of the fact that the gamma random effects have unit mean.

3 Generation of correlated counts

As will be presented in Section 3.1, the GLMM can be used to parsimoniously generate correlated count data with prespecified marginal mean function and such variance-covariance structures as compound symmetry and the one generated by random intercept and random slope. In the GLMM

case, however, the random effects used do not separate correlation and overdispersion, a disadvantage that may lead to mis-representation of the random-effects variability. The algorithm for generating data from the combined model, which accounts for both correlation and overdispersion, is given in Section 3.2.

3.1 The GLMM as a data generator

The GLMM can be used to generate correlated random variables with a desired structure. Given a marginal (log) mean (possibly depending on covariates $\widetilde{X}_i$) and a variance-covariance matrix for $\mathbf{Y}_i$, Algorithm 1 below generates random variables with this pre-specified structure.

Algorithm 1:

1. *Derive the unknowns $\boldsymbol{\beta}$ and D of the GLMM by comparing the desired marginals with the marginals from the GLMM.*
2. *Using D, simulate $\mathbf{b}_i$.*
3. *Compute* $\ln(\lambda_{ij}) = \boldsymbol{x}_{ij}^\top\boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top\mathbf{b}_i$.
4. *Simulate* $Y_{ij} \sim Poi(\lambda_{ij})$.

To put matters into context, if we consider the case of compound symmetry (CS), for example, in that the desired marginal mean is $\ln(\boldsymbol{\mu}_i) = \widetilde{X}_i\alpha$ and desired variance-covariance structure is $V = M_i + \tau^2 J_i$ (CS structure), then the necessary unknowns in step 1 of the above algorithm are derived by comparing [a] $\widetilde{X}_i\alpha = X_i\boldsymbol{\beta} + 0.5Z_iDZ_i^\top$ [which is (2a) expressed in matrix form] for the marginal mean, and, [b] $M_i + \tau^2 J_i = M_i + M_i(e^{Z_iDZ_i^\top} - J_i)M_i$ for the marginal variance-covariance structure. Solving [a] for $\boldsymbol{\beta}$ and [b] for D leads to:

$$\boldsymbol{\beta} = (X_i^\top X_i)^- X_i^\top(\widetilde{X}_i\alpha - 0.5Z_iDZ_i^\top), \tag{6a}$$

$$D = (Z_i^\top Z_i)^- Z_i^\top \log\left(M_i^{-1}\tau^2 J_i M_i^{-1} + J_i\right) Z_i (Z_i^\top Z_i)^-, \tag{6b}$$

where $(.)^-$ indicates a generalized inverse. For a general V, $\tau^2 J_i$ in D above becomes $V - M_i$. Then, it follows that $E(\mathbf{Y}_i) = e^{\widetilde{X}_i\alpha}$ and $\text{var}(\mathbf{Y}_i) = V$. If the generalized inverse is not an inverse, the solution clearly is not unique. This is not a problem, it simply means that several choices of $\boldsymbol{\beta}$ and D are possible, that nevertheless all lead to the desired marginal structure. This is akin to the fact that there is a one-to-many map between a given marginal model on the one hand and the class of hierarchical models that marginalizes to it on the other. Any member of the class of hierarchical model can in principle be used as a data generator for the marginal structure.

3.2 The combined model as a data generator

The combined model can be used to generate correlated Poisson random variables following logic similar to that described in Section 3.1. The major difference from the GLMM is that there is a third unknown term in the combined model, i.e., Σ_i, the variance-covariance matrix for the overdispersion parameter(s). Given a desired mean and variance-covariance structure, Algorithm 2 generates the Poisson variates.

Algorithm 2:

1. *Derive the unknowns $\boldsymbol{\beta}$, D, and Σ_i in the CM.*
2. *Generate $\boldsymbol{\theta}_i \sim MGamma(mean = 1, variance = \Sigma_i)$.*
3. *Using D, simulate $\mathbf{b}_i$.*
4. *Compute $\lambda^*_{ij} = \theta_{ij} \exp(x_{ij}^\top \boldsymbol{\beta} + z_{ij}^\top \mathbf{b}_i)$.*
5. *Simulate $Y_{ij} \sim Poi(\lambda^*_{ij})$.*

The necessary unknowns in step 1 of Algorithm 2 are given by $\boldsymbol{\beta}$ as in (6a) and further

$$D = (Z_i^\top Z_i)^- Z_i^\top \log\left[M_i^{-1}(V - M_i)M_i^{-1} + J_i\right] Z_i (Z_i^\top Z_i)^-,$$

$$\Sigma_i = e^{-Z_i D Z_i^\top}\left[M_i^{-1}(V - M_i)M_i^{-1} + J_i\right] - J_i,$$

where notational conventions are as before.

An extension to generating purely serially correlated outcomes is done by removing the normal random effect and choosing $\boldsymbol{\theta}_i$ such that it follows a serially correlated multivariate gamma. Note that 'multivariate' is used here in the broad sense, because all hierarchical structures, such as longitudinal and clustered data to name a few, imply marginal multivariate structures. Evidently, in such structured designs, the marginal covariance matrix will typically not be unstructured.

The general form of the combined model (4), in the case of Poisson data, is that the normal random effects are correlated and the Gamma random effects are also correlated. From this general case, several special cases can be derived. An overview of the possible combinations is presented in Table 1. The following special cases, which are also presented in Table 1, can be derived from the more general case:

- A combination of normal and independent Gamma random effects. This is the most commonly used form of the combined model in which the normal random effects induce/account for correlation while the Gamma random effects induce/account for overdispersion. It is model (4) but with Σ_i diagonal.
- Normal random effects without Gamma random effects. In this case, (4) reduces to (1) and data is generated as explained in Section 3.1. Here, the normal random effects induce/account for both correlation and overdispersion.
- No normal random effects, no Gamma random effects. The absence of both random effects is equivalent to generating independent counts which is not of interest in this paper.
- No normal random effects, correlated Gamma random effects such that both correlation and overdispersion are induced via the Gamma random effects. Thus, λ_{ij} in (4) becomes $\exp(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta})$ and Σ_i is fully general.
- No normal random effects, independent Gamma random effects. In this case, the combined

model reduces to the negative-Binomial model which accounts for overdispersion but not correlation. λ_{ij} in (4) becomes $\exp(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta})$ and Σ_i is diagonal.

Extra variations can be constructed by choosing for the normal random effects (random intercept + slope, or higher dimensions) to be either independent (D diagonal) or correlated. In this paper, we have only studied the latter case but the former is very easily obtainable.

4 Setup of simulation study

As illustrated in Section 3.2, the combined model can take on several forms or variations. To evaluate the performance of the different forms of the combined model as data generators, a simulation was set up across the variations. More specifically, given a pre-specified marginal mean and variance-covariance matrix, 1000 Monte Carlo replications of correlated count data sets were generated from each of the several forms. Marginal models were then fitted to these data sets and the difference between the pre-specified parameters and those estimated by fitting the marginal models were studied. Two different arms have been considered for the simulation, namely, sample size $K = 100$ and 500. For $K = 100$, 2 correlated Poisson variables were generated from the following model specification;

$$
\begin{aligned}
Y_{ij} &\sim \text{Poi}(\lambda_{ij}^{*}),\\
\lambda_{ij}^{*} = \theta_{ij}\lambda_{ij} &= \theta_{ij}\exp(\beta_0 + b_{0i} + \beta_1 T_i + (\beta_2 + b_{1i})t_{ij} + \beta_3 T_i * t_{ij}),\\
\boldsymbol{\theta}_i &\sim \text{Gamma}(\text{mean} = \mathbf{1}, \text{variance} = \Sigma_i),\\
\mathbf{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, D = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}\right],\\
V^{*} &= \begin{pmatrix} 36 & 12 \\ 12 & 29 \end{pmatrix},
\end{aligned}
\tag{7}
$$

where $T_i \sim$ Bernoulli(0.5), t_{ij} is the ordering of the j^{th} observation ($i = 1, \ldots, K = 100$, $j = 1, 2$) in subject i, and the desired marginal mean parameters are $\beta_0 = 1.521, \beta_1 = 0.237, \beta_2 = 0.254, \beta_3 = 0.345$. Generalized estimating equations (GEE, Liang and Zeger 1986), NEGBIN, and the GLMM were used to study the behavior of the data generator, averaged over the 1000 MC replications. GEE is one tool commonly used to model correlated data when scientific interest is in inference on the marginal parameters. It makes no distributional assumptions apart from the specification of the mean function $\boldsymbol{\mu}_i = \exp(\mathbf{X}_i\boldsymbol{\beta})$ for models with the log link, the variance function $V_i = A_i^{1/2} R_i(\boldsymbol{\alpha}) A_i^{1/2}$ where A_i is an $n_i \times n_i$ diagonal matrix with var(μ_{ij}) as the j^{th} diagonal element, and $R_i(\boldsymbol{\alpha})$ is an $n_i \times n_i$ (perhaps incorrect) working correlation matrix to allow for dependence between within-subject observations expressed in terms of $\boldsymbol{\alpha}$ a vector of unknown parameters.

For $K = 500$, 4 random variables were generated from a similar model as above, the difference being that a random intercept model for the normal random effects was used. More specifically,

$$\begin{aligned} \lambda^*_{ij} &= \theta_{ij}\lambda_{ij} = \theta_{ij}\exp(\beta_0 + b_{0i} + \beta_1 T_i + \beta_2 t_{ij} + \beta_3 T_i * t_{ij}), \\ \mathbf{b}_i &= b_{0i} \sim N(0, d), \\ V^* &= \begin{pmatrix} 256 & 128 & 144 & 224 \\ 128 & 208 & 228 & 172 \\ 144 & 228 & 299 & 296 \\ 224 & 172 & 296 & 567 \end{pmatrix}, \end{aligned} \tag{8}$$

where $i = 1, \ldots, K = 500$ and $j = 1, 2, 3, 4$. The desired marginal mean parameters were specified as $\beta_0 = 1.521, \beta_1 = 0.437, \beta_2 = -0.254$ and $\beta_3 = 0.145$. In addition to GEE, NEGBIN and GLMM models used in the case of $K = 100$, the so-called marginal multilevel model (MMM) was also used, mainly motivated by the fact that the sensitivity of the MMM to starting values is less severe if the random intercept model is specified for the normal random effects than in the case of random intercept and slope. The MMM was described by Heagerty (1999) for binary longitudinal data,

building on a specification of the marginal rather than the conditional mean given random effects. More precisely, this model puts together the two worlds of marginal and conditional or hierarchical modeling in the sense that it puts the ideas of Generalized Estimating Equations (GEE, Liang and Zeger, 1986) and the GLMM together leading to inferences both in the marginal and conditional senses.

5 Results of Simulation Study

Tables 2 and 3 present the results from the simulation study. Generally, from Table 2, all marginal models (GEE, NEGBIN, MMM) seem to perform similarly across the various forms of the combined model. This is expected as the proposed data generator is aimed at the context of marginal models. Specific to this case of using a random-intercept model for the normal random effects, GEE, MMM, and the GLMM yield the same results for time-related parameters β_2 and β_3 with minor differences between GEE or GLMM versus MMM in the case of normal and no gamma random effects. Given normal random effects with random intercept only and no Gamma random effects, the marginal parameters are expected to be the same as the hierarchical parameters with a change in β_0. Indeed, GEE, NEGBIN, MMM, and GLMM yield the same parameter estimates with a change in the intercept (β_0) for GLMM. Across all variations of the combined model, GEE, MMM, and GLMM generally differ on parameters β_0 and β_1. No specific pattern can be identified for the NEGBIN relative to GEE and MMM, except in the above-mentioned case of normal random effects and no Gamma random effects. When the Gamma random effects are correlated, the parameter estimates are rather different from the true parameters and even change sign for β_2 for both hierarchical and marginal models. Since the GLMM is a hierarchical model, the results for the GLMM presented should be interpreted with caution. We emphasize that GLMM should not be used to model data generated by our proposal. From Table 3, which is the case of a random

intercept and slope model for the normal random effects, both GEE and NEGBIN yield the same parameter estimates and standard deviations across the combined model variations. Since in this setting, only 2 random variables were generated, it may be interesting to consider the generation of more than 2 random variables and also larger sample sizes so as to get broader insight into this scenario. The parameter α for the NEGBIN goes to infinity in the absence of overdispersion, which is what we observe in the normal RE, no Gamma RE case. Again, the GLMM should be interpreted with care given that it is not a marginal but rather a hierarchical model.

Apart from the simulation, we also generated 4 different datasets of size $K = 500$ from the combined model with [1] two time points (bivariate case) with only the random intercept specified for the $\mathbf{b}_i$ random effects, [2] two time points with random intercept and slope, [3] four time points with random intercept only, and [4] four time points with random intercept and slope. The gamma random effects are correlated. Table 4 summarizes the generation settings considered here, in which 2 or 4 correlated Poisson variates are generated corresponding to 2 and 4 time points, respectively. We have only considered the case of the random intercept on the one hand and the random intercept and slope in time models on the other, for illustrative purposes. It is easy to manipulate more general dimensions. Note though that the higher the random-effects dimension, the higher the risk of the D matrix not being positive-definite. Also, because the gamma random effects are allowed to be correlated, very little or no information is derived from the $\mathbf{b}_i$ random effects. We generate data given covariates ($\widetilde{X}_i$) as treatment (trt, 0 or 1), time (2 or 4 points) and the interaction of treatment and time. Note that we assume $\widetilde{X}_i = X_i$, thus using the same covariates but the method also allows for use of different covariates in the two design matrices. Table 5 shows the results of the derived unknown parameters that aid the data generation process for the 4 cases presented in Table 4. Here, $\boldsymbol{\alpha}$ is the parameter vector for the specified marginal mean and *diff* is the change between the marginal parameters $\boldsymbol{\alpha}$ and the conditional/derived parameters $\boldsymbol{\beta}$. As expected in the case of a random intercept model (cases 1 and 3), a change is only evident in the intercept

relative to the other parameters. In cases 2 and 4 for the random intercept and slope model, a difference between the marginal and conditional mean parameters is reflected in the intercept and time parameter estimates. Table 6 presents the summary statistics and the Spearman correlation coefficients of the generated Poisson variables, while Figures 1–4 show marginal distributions and scatter plots of the generated random variables for cases 1–4, respectively. In Table 6, the mean is smaller than the square of the standard deviation, indicating overdispersion. It can also be seen that the generated random variables are correlated (see ρ). From Table 4, cases 3 and 4 are similar with the only difference being that case 3 only has a random intercept while case 4 has random intercept and slope(time) as the covariates for the random effects. Specific to this case and given that Σ_i is fully general, there are minimal changes from case 3 to 4 (see Figures 3 and 4, and Table 6). Similarly, by comparing Figures 1 and 2, and also Table 6, we clearly see that that inclusion of a random slope allows to roughly retain the correlation structure, but modifies the mean and variance structures. Further, when the marginal structure is specified, it is possible to decompose the hierarchical structure (in particular, the random effects) in different ways, yet leading the same result, as it should be. Indeed, it is clear, from comparing Figures 3 and 4, that the same marginal structure (mean, variance, correlation) can be obtained, with our without the use of a random slope. This gives the user some latitude as to choose a decomposition that is flexible yet computationally efficient.

6 Discussion and conclusions

The combined model as introduced by Molenberghs et al. (2007, 2010) simultaneously accommodates correlation and overdispersion unexplained by the normal random effects, In the absence of correlation, the model simplifies to a negative-binomial model for overdispersion. On the other hand, in the absence of overdispersion, it simplifies to the GLMM. The model's flexible capabili-

ties make it a good candidate as a data generator given that one always wants to generate data that reflects the characteristics of interest, in this case, overdispersion and/or correlation. The CM is a convenient tool that mimics or incorporates these intrinsic features of correlated count data. In particular, a fully marginal view as well as a random-effects view can be taken. This implies that a broad toolkit emerges. In the purely marginal view, essentially a multivariate gamma variate, easy to generate, is transformed to a multivariate count variable.

The covariates determining the fixed- and random-effects design matrices are kept simple herein. This is not limiting in the sense that a specification of any covariates can be done as is needed. It is possible to encounter non-positive definite D matrices or negative entries along the diagonal of Σ_i. This may point to a non-allowable hierarchical model to come with the marginal model or perhaps a marginal model that is in itself not allowable. The analogy would be a multivariate normal with a given but non-positive definite variance-covariance matrix. Such model is invalid in the first place and needs to be reconsidered.

Because the combined model is hierarchical, random variables with only positive correlations are generated due to restrictions of positive-definiteness on the random effects variance-covariance matrices. This may be a drawback for the combined model, as is the case for some of the methods present in literature for count data generation. However, a way to overcome this is to generate directly from the marginal model, arguably via correlated θ_{ij}, of which the variance-covariance matrix Σ_i then reflects the desired structure.

Acknowledgments

The authors gratefully acknowledge support from IAP research Network P7/06 of the Belgian Government (Belgian Science Policy).

References

Agresti, A. (2000). *Categorical Data Analysis (2nd ed.)*. New York: John Wiley & Sons.

Avramidis, A.N., Channouf, N. and L'Ecuyer, P. (2009). Efficient correlation matching for fitting discrete multivariate distributions with arbitrary marginals and normal-copula dependence. *INFORMS Journal on Computing* **2**, 88–106.

Breslow, N. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics* **33**, 38–44.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Booth, J.G., Casella, G., Friedl, H., and Hobert, J.P. (2003). Negative binomial loglinear mixed models. *Statistical Modelling* **3**, 179–181.

Cario, M.C. and Nelson, B.L. (1997). Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Tech. rep., Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.

Cario, M.C. and Nelson, B.L. (1998). Numerical methods for fitting and simulating autoregressive-to-anything processes. *INFORMS Journal on Computing* **10**, 72–81.

Chin, H.C. and Quddus, M.A. (2003). Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis & Prevention* **35**, 253–259.

Deb, P. and Holmes, A. M. (2000). Estimates of use and costs of behavioural health care: A comparison of standard and finite mixture models. *Health Economics* **9**, 475–489.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York: Springer.

Downer, R. and Moser, E. (2001). On the generation of a multivariate spatial poisson distribution, Lousiana State University, Department of Experimental Statistics, Technical Reports RR-O1–35.

Ghosh, S. and Pasupathy, R. (2012). C-NORTA: A Rejection Procedure for Sampling from the Tail of Bivariate NORTA Distributions. *INFORMS Journal on Computing* **24**, 295–310.

Heagerty, P.J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* **55**, 688–698.

Hinde, J. and Demétrio, C.G.B. (1998a). Overdispersion: Models and estimation. *Computational Statistics and Data Analysis* **27**, 151-170.

Hinde, J. and Demétrio, C.G.B. (1998b). *Overdispersion: Models and Estimation*. São Paulo: XIII Sinape.

Karlis, D. (2003). An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics* **30**, 63–77.

Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions*. Boca Raton: CRC Press.

Kocherlakota, S. and Kocherlakota, K., (2001). Regression in the bivariate Poisson distribution. *Communications in Statistics, Theory & Methods* **30**, 815–825.

Lawless, J.F. (1987). Negative Binomial and Mixed Poisson regression. *The Canadian Journal of Statistics* **15**, 209–225.

Li, S.T. and Hammond, J.L. (1975). Generation of pseudo-random numbers with specified univariate distributions and correlation coefficients. *IEEE Transactions on Systems, Man, and Cybernetics* **5**, 557–561.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Madsen, L. and Dalthorp, D. (2007). Simulating correlated count data. *Environmental and Ecological Statistics* **14**, 129–148.

Mardia, K.V. (1970). *Families of Bivariate Distributions*. London: Griffin.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman & Hall.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

Molenberghs, G., Verbeke, G., and Demétrio, C. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis* **13**, 513–531.

Molenberghs, G., Verbeke, G., Demétrio, C., and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science* **25**, 325–347.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series B*, **135**, 370–384.

Nelsen, R.B. (2006). *An Introduction to Copulas*. Berlin: Springer.

Park, C.G. and Shin, D.W. (1998). An algorithm for generating correlated random variables in a class of infinitely divisible distributions. *Journal of Statistical Computation and Simulation* **61**, 127–139.

Shin, K. and Pasupathy, R. (2010). An algorithm for fast generation of bivariate Poisson random vectors. *INFORMS Journal on Computing* **22**, 81–92.

Sutradhar, B. (2011). *Dynamic Mixed Models for Familial Longitudinal Data*. New York: Springer.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. Berlin: Springer-Verlag.

Winkelmann, R. (2004). Health care reform and the number of doctor visits. An econometric analysis. *Journal of Applied Econometrics* **19**, 455–472.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computing and Simulation* **48**, 233–243.

Yahav, I. and Shmueli, G. (2012). On generating multivariate Poisson data in management science applications. *Applied Stochastic Models for Business and Industry*, **28**, 91–102.

Table 1: *Possible combinations of the normal and Gamma random effects in the context of count data. ✓ refers to combinations of the combined model from which correlated and/or overdispersed data can be generated, while ✗ refers to the independent count data generation case*

			Gamma random effects		
	Present		Yes		No
			Correlated	Independent	
Normal random effects	Yes	Correlated	✓	✓	✓
		Independent	✓	✓	✓
	No		✓	✓	✗

Table 2: *Simulation, generate 4 random variables: Parameter estimates (standard deviations) for GEE (exchangeable correlation), NEGBIN, MMM and GLMM, and, absolute bias (MSE) for GEE, NEGBIN and MMM, averaged over 1000 MC replications for sample size (N) = 500. True parameters are* $\beta_0 = 1.521, \beta_1 = 0.437, \beta_2 = -0.254$ *and* $\beta_3 = 0.145$ *and a random intercept model was specified for the normal random effects (RE). Corr means correlated while IND means independent*

Model	Parameter	Normal, Corr Gamma	Normal, IND Gamma	Normal, No Gamma	No normal, Corr Gamma	No normal, IND Gamma
Parameter estimates (standard deviations)						
GEE	intercept (β_0)	3.353(0.098)	1.550(0.515)	1.522(0.046)	3.354(0.098)	1.556(0.509)
	T (β_1)	1.295(0.103)	0.413(0.564)	0.437(0.057)	1.295(0.103)	0.408(0.561)
	t (β_2)	0.015(0.040)	-0.296(0.259)	-0.255(0.018)	0.015(0.040)	-0.298(0.256)
	t*T (β_3)	0.079(0.041)	0.178(0.281)	0.145(0.022)	0.079(0.041)	0.180(0.279)
NEGBIN	intercept (β_0)	3.319(0.185)	1.681(0.633)	1.522(0.046)	3.321(0.185)	1.690(0.629)
	T (β_1)	0.978(0.193)	0.296(0.678)	0.437(0.057)	0.979(0.194)	0.289(0.677)
	t (β_2)	0.028(0.085)	-0.354(0.320)	-0.255(0.018)	0.027(0.085)	-0.358(0.318)
	t*T (β_3)	0.203(0.088)	0.231(0.341)	0.145(0.022)	0.202(0.088)	0.234(0.340)
	alpha	0.322(0.009)	0.064(0.004)	899.005(1305.540)	0.322(0.010)	0.064(0.004)
MMM	intercept (β_0)	3.066(0.118)	1.395(0.617)	1.522(0.050)	3.066(0.118)	1.404(0.608)
	T (β_1)	2.174(0.142)	2.038(0.669)	0.431(0.062)	2.178(0.141)	2.040(0.664)
	t (β_2)	0.015(0.040)	-0.296(0.259)	-0.255(0.022)	0.015(0.040)	-0.298(0.256)
	t*T (β_3)	0.079(0.041)	0.178(0.281)	0.146(0.022)	0.079(0.041)	0.180(0.279)
	d	1.448(0.120)	5.630(0.447)	0.006(0.004)	1.451(0.121)	5.660(0.454)
GLMM	intercept (β_0)	2.342(0.136)	-1.421(0.620)	1.519(0.045)	2.341(0.135)	-1.426(0.612)
	T (β_1)	2.174(0.142)	2.038(0.669)	0.437(0.055)	2.178(0.141)	2.040(0.664)
	t (β_2)	0.015(0.040)	-0.296(0.259)	-0.255(0.018)	0.015(0.040)	-0.298(0.256)
	t*T (β_3)	0.079(0.041)	0.178(0.281)	0.145(0.022)	0.079(0.041)	0.180(0.279)
	d	1.448(0.120)	5.630(0.447)	0.006(0.004)	1.451(0.121)	5.660(0.454)
Absolute bias (MSE)						
GEE	intercept (β_0)	1.832(3.365)	0.029(0.266)	0.001(0.002)	1.833(3.369)	0.035(0.261)
	T (β_1)	0.858(0.747)	0.024(0.318)	0.000(0.003)	0.858(0.746)	0.029(0.316)
	t (β_2)	0.269(0.074)	0.042(0.069)	0.001(0.000)	0.269(0.074)	0.044(0.067)
	t*T (β_3)	0.066(0.006)	0.033(0.080)	0.000(0.000)	0.066(0.006)	0.035(0.079)
NEGBIN	intercept (β_0)	1.798(3.268)	0.160(0.426)	0.001(0.002)	1.800(3.273)	0.169(0.424)
	T (β_1)	0.541(0.331)	0.141(0.479)	0.000(0.003)	0.542(0.331)	0.148(0.481)
	t (β_2)	0.282(0.087)	0.100(0.113)	0.001(0.000)	0.281(0.086)	0.104(0.112)
	t*T (β_3)	0.058(0.011)	0.086(0.124)	0.000(0.000)	0.057(0.011)	0.089(0.124)
MMM	intercept (β_0)	1.545(2.401)	0.126(0.396)	0.001(0.002)	1.545(2.402)	0.117(0.383)
	T (β_1)	1.737(3.037)	1.601(3.011)	0.006(0.004)	1.741(3.051)	1.603(3.008)
	t (β_2)	0.269(0.074)	0.042(0.069)	0.001(0.000)	0.269(0.074)	0.044(0.067)
	t*T (β_3)	0.066(0.006)	0.033(0.080)	0.001(0.001)	0.066(0.006)	0.035(0.079)

Table 3: *Simulation, generate 2 random variables: Parameter estimates (standard deviations) for GEE (exchangeable correlation), NEGBIN and GLMM, and, absolute bias (MSE) for GEE and the NEGBIN models averaged over 1000 MC replications, N=100. True parameters are* $\beta_0 = 1.521, \beta_1 = 0.237, \beta_2 = 0.254, \beta_3 = 0.345$ *with a random intercept and slope model specified for the normal random effects (RE). Corr means correlated while IND means independent*

Model	Parameter	Normal, Corr Gamma	Normal, IND Gamma	Normal, No Gamma	No normal, Corr Gamma	No normal, IND Gamma
Parameter estimates (standard deviations)						
GEE	intercept (β_0)	1.749(0.240)	1.522(0.304)	1.524(0.127)	1.752(0.238)	1.497(0.304)
	T (β_1)	0.922(0.258)	0.226(0.350)	0.241(0.157)	0.866(0.260)	0.253(0.351)
	t (β_2)	0.266(0.127)	0.249(0.175)	0.252(0.078)	0.263(0.126)	0.264(0.175)
	t*T (β_3)	0.336(0.136)	0.355(0.198)	0.342(0.095)	0.337(0.137)	0.338(0.199)
NEGBIN	intercept (β_0)	1.749(0.240)	1.522(0.304)	1.524(0.127)	1.752(0.238)	1.497(0.304)
	T (β_1)	0.922(0.258)	0.226(0.350)	0.241(0.156)	0.866(0.260)	0.253(0.351)
	t (β_2)	0.266(0.127)	0.249(0.175)	0.253(0.078)	0.263(0.126)	0.264(0.175)
	t*T (β_3)	0.336(0.136)	0.355(0.198)	0.342(0.095)	0.337(0.137)	0.338(0.199)
	alpha	6.494(1.628)	3.715(0.729)	1238.877(1640.875)	6.316(1.492)	3.740(0.714)
GLMM	intercept (β_0)	1.373(0.244)	1.003(0.311)	1.508(0.066)	1.373(0.240)	0.988(0.311)
	T (β_1)	1.186(0.263)	0.484(0.353)	0.218(0.082)	1.129(0.261)	0.502(0.353)
	t (β_2)	0.427(0.127)	0.464(0.177)	0.242(0.044)	0.425(0.126)	0.474(0.176)
	t*T (β_3)	0.223(0.136)	0.252(0.198)	0.383(0.057)	0.225(0.136)	0.240(0.199)
	d_{11}	0.806(0.224)	1.997(0.457)	0.043(0.069)	0.824(0.219)	1.980(0.446)
	d_{12}	-0.339(0.105)	-1.058(0.255)	-0.031(0.039)	-0.347(0.103)	-1.051(0.246)
	d_{22}	0.155(0.052)	0.590(0.147)	0.225(0.141)	0.159(0.052)	0.588(0.140)
Absolute bias (MSE)						
GEE	intercept (β_0)	0.228(0.110)	0.001(0.093)	0.003(0.016)	0.231(0.110)	0.024(0.093)
	T (β_1)	0.685(0.536)	0.011(0.123)	0.004(0.025)	0.629(0.463)	0.016(0.124)
	t (β_2)	0.012(0.016)	0.005(0.031)	0.002(0.006)	0.009(0.016)	0.010(0.031)
	t*T (β_3)	0.009(0.018)	0.010(0.039)	0.003(0.009)	0.008(0.019)	0.007(0.040)
NEGBIN	intercept (β_0)	0.228(0.110)	0.001(0.093)	0.003(0.016)	0.231(0.110)	0.024(0.093)
	T (β_1)	0.685(0.536)	0.011(0.123)	0.004(0.025)	0.629(0.463)	0.016(0.124)
	t (β_2)	0.012(0.016)	0.005(0.031)	0.001(0.006)	0.009(0.016)	0.010(0.031)
	t*T (β_3)	0.009(0.018)	0.010(0.039)	0.003(0.009)	0.008(0.019)	0.007(0.040)

Table 4: *Parameters specified to generate correlated Poisson random variables from the combined model.*

	2 time points	4 time points
	Case 1:	Case 3:
$\widetilde{X}_i = X_i$ covariates	*Intercept T t T*t*	*Intercept T t T*t*
α	1.521 0.237 0.254 0.345	1.521 0.437 -0.254 0.145
Z_i covariates	*Intercept*	*Intercept*
V^*	$\begin{bmatrix} 36 & 12 \\ 12 & 29 \end{bmatrix}$	$\begin{bmatrix} 256 & 128 & 144 & 224 \\ 128 & 208 & 228 & 172 \\ 144 & 228 & 299 & 296 \\ 224 & 172 & 296 & 567 \end{bmatrix}$
	Case 2:	Case 4:
covariates ($\widetilde{X}_i = X_i$)	*Intercept T t T*t*	*Intercept T t T*t*
α	2.521 0.237 0.254 0.345	1.521 0.437 -0.254 0.145
Z_i covariates	*Intercept + t*	*Intercept + t*
V^*	$\begin{bmatrix} 225 & 615 \\ 615 & 2581 \end{bmatrix}$	$\begin{bmatrix} 256 & 128 & 144 & 224 \\ 128 & 208 & 228 & 172 \\ 144 & 228 & 299 & 296 \\ 224 & 172 & 296 & 567 \end{bmatrix}$

Table 5: *The necessary unknowns (β and D) for each of the cases presented in Table 4.*

		Derived unknowns			
Case	*Parameter*	α	β	*diff*	*D*
1.	Intercept	1.521	1.521	0.0002203	$\begin{bmatrix} 0.0004406 \end{bmatrix}$
	T	0.237	0.237	-1.02E-14	
	t	0.254	0.254	-6.88E-15	
	T*t	0.345	0.345	1.082E-14	
2.	Intercept	2.521	2.521	-0.000493	$\begin{bmatrix} 0.000263 & -0.000039 \\ -0.000039 & 0.0006242 \end{bmatrix}$
	T	0.237	0.237	6.495E-15	
	t	0.254	0.253	0.0008976	
	T*t	0.345	0.345	-3.89E-15	
3.	Intercept	1.521	1.518	0.002885	$\begin{bmatrix} 0.00577 \end{bmatrix}$
	T	0.437	0.437	-3.4E-14	
	t	-0.254	-0.254	-1.11E-15	
	T*t	0.145	0.145	-1.5E-15	
4.	Intercept	1.521	1.520	0.0014135	$\begin{bmatrix} 0.0040014 & 0.0000601 \\ 0.0000601 & 0.0002349 \end{bmatrix}$
	T	0.437	0.437	-3.5E-14	
	t	-0.254	-0.255	0.0006473	
	T*t	0.145	0.145	6.939E-16	

Table 6: *Summary statistics and the Spearman correlation (ρ) matrices of the generated Poisson variables; std refers to the standard deviation.*

Case	Var.	mean	std	median	min.	max.	ρ
1.	Y1	16.68	12.15	15.00	0	56	$\begin{bmatrix} 1 & 0.81 \\ & 1 \end{bmatrix}$
	Y2	28.36	19.93	27.50	0	75	
2.	Y1	87.07	43.92	85.50	7	205	$\begin{bmatrix} 1 & 0.88 \\ & 1 \end{bmatrix}$
	Y2	142.86	121.50	121.50	0	654	
3.	Y1	16.39	29.35	3	0	171	$\begin{bmatrix} 1 & 0.64 & 0.60 & 0.59 \\ & 1 & 0.96 & 0.69 \\ & & 1 & 0.79 \\ & & & 1 \end{bmatrix}$
	Y2	129.10	106.93	110	0	498	
	Y3	152.90	141.64	127	0	619	
	Y4	27.26	72.56	0	0	832	
4.	Y1	16.71	30.50	3.50	0	184	$\begin{bmatrix} 1 & 0.68 & 0.64 & 0.60 \\ & 1 & 0.96 & 0.68 \\ & & 1 & 0.79 \\ & & & 1 \end{bmatrix}$
	Y2	129.94	109.32	113.00	0	601	
	Y3	155.41	147.01	127.50	0	748	
	Y4	28.84	84.72	0	0	1147	

```
        Generate Correlated Poisson data from CM with
          Normal and Correlated Gamma random effects
*********************************************************

                  Sample size (K) = 500
        minimum number of measurements per subject = 4
        maximum number of measurements per subject = 4
         Normal random effects covariates = Intercept


      Given Mean parameters are: Intercept =     1.521
                                 trt0            0.437
                                 time           -0.254
                                 trt0time        0.145

                                          Y1        Y2        Y3        Y4

Given variance-covariance matrix =  Y1       256       128       144       224
                                    Y2       128       208       228       172
                                    Y3       144       228       299       296
                                    Y4       224       172       296       567


     Parameter      alpha       beta       diff          D

     Intercept      1.521   1.518115   0.002885    0.00577
     trt0           0.437      0.437 -3.13E-14
     time          -0.254     -0.254  1.61E-15
     trt0time       0.145      0.145 -3.47E-15
```

Figure 1: *Two Poisson random variables generated from the combined model with random intercept model.*

```
      Generate Correlated Poisson data from CM with
            Normal and No Gamma random effects
  ************************************************

                Sample size (K) = 20
      minimum number of measurements per subject = 2
      maximum number of measurements per subject = 2
       Normal random effects covariates = Intercept


    Given Mean parameters are: Intercept =     1.521
                               trt0            0.237
                               time           -0.254
                               trt0time       -0.345

                                                 Y1         Y2

Given variance-covariance matrix =  Y1       0.49       0.63
                                    Y2       0.63       4.81


    Parameter      alpha       beta       diff          D

    Intercept      1.521  1.5190213  0.0019787  0.0039574
    trt0           0.237      0.237 -5.88E-15
    time          -0.254     -0.254    -4E-15
    trt0time      -0.345     -0.345  5.052E-15
```

Figure 2: *Two Poisson random variables generated from the combined model with random intercept and slope model.*

Generate Correlated Poisson data from CM with
Correlated Normal and Correlated Gamma random effects
****

Sample size (K) = 500
minimum number of measurements per subject = 4
maximum number of measurements per subject = 4
Normal random effects covariates = Intercept time

Given Mean parameters are: Intercept = 1.521
trt0 0.437
time -0.254
trt0time 0.145

Given variance-covariance matrix =

	Y1	Y2	Y3	Y4
Y1	256	128	144	224
Y2	128	208	228	172
Y3	144	228	299	296
Y4	224	172	296	567

Parameter	alpha	beta	diff	D	
Intercept	1.521	1.5195865	0.0014135	0.0040014	0.0000601
trt0	0.437	0.437	-1.59E-14	0.0000601	0.0002349
time	-0.254	-0.254647	0.0006473		
trt0time	0.145	0.145	-1.86E-15		

Figure 3: *Four Poisson random variables generated from the combined model with random intercept model.*

```
   Generate Correlated Poisson data from CM with
Correlated Normal and Correlated Gamma random effects
 ****************************************************

                Sample size (K) = 500
   minimum number of measurements per subject = 2
   maximum number of measurements per subject = 4
 Normal random effects covariates = Intercept time


  Given Mean parameters are: Intercept =     1.521
                             trt0            0.437
                             time           -0.254
                             trt0time        0.145

                                        Y1        Y2        Y3        Y4

Given variance-covariance matrix =  Y1       256       128       144       224
                                    Y2       128       208       228       172
                                    Y3       144       228       299       296
                                    Y4       224       172       296       567


      Parameter     alpha      beta      diff         D

      Intercept     1.521 1.5209283 0.0000717 0.0071187 -0.001993
      trt0          0.437 0.4370443 -0.000044 -0.001993 0.0016015
      time         -0.254 -0.255718 0.0017181
      trt0time      0.145 0.1449747 0.0000253
```

Figure 4: *Four Poisson random variables generated from the combined model with random intercept and slope model.*